DOCUMENT RESUME

ED 079 384 TM 002 984

AUTHOR	Doppelt, Jermoe E.
TITLE	Watch Your Weights.
INSTITUTION	Psychological Corp., New York, N.Y.
PUB DATE	57
NOTE	4p.; Reprint from previous Test Service Bulletin
JOURNAL CIT	Test Service Bulletin, n52-52 p2-5 1957-1958
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Bulletins; *Scoring; Statistical Analysis; *Test Interpretation; Test Results; *Weighted Scores

ABSTRACT

The toal score concept is considered from the test user's point of views. The simples kind of total, obtained by adding raw scores on the tests, automatically assigns weights to the tests which are proportional to their standard deviation. When desired or appropriate, tests can be weighted equally in a total by transforming the raw scores on each test so that standard deviations will be the same. This procedures, usually done by converting the scores on each test to scaled scores, does not of itself provide a better measuring instrument. Validation studies must still be conducted. For a particular criterion, the most predictive type of total score is one based on optimum weights for the tests. In some instances, test authors have so constructed their tests that a simple summation of scores is also the best-weighted total for a specific type of criterion. It is not safe to assume this however; the data should be examined carefully. (Author/KM)

ED 079384

Test Service Bulletin

No. 52 THE PSYCHOLOGICAL CORPORATION December, 1957

Published from time to time in the interest of promoting greater understanding of the principles and techniques of mental measurement and its applications in guidance, personnel work, and clinical psychology, and for announcing new publications of interest. Address communications to 304 East 45th Street, New York 17, N. Y.

HAROLD G. SEASHORE, *Editor*
Director of the Test Division

ALEXANDER G. WESMAN
Associate Director of the Test Division

JEROME E. DOPPELT
Assistant Director

JAMES H. RICKS, JR.
Assistant Director

DOROTHY M. CLENDENEN
Assistant Director

ESTHER R. HOLLIS
Advisory Service

WATCH YOUR WEIGHTS

THE work of the personnel man, counselor, or psychologist would be very simple if he needed just one bit of information about each person for each decision to be made. It would be so simple, in fact, that the professional person who makes the decision could quickly be replaced by a machine. What complicates his life, slows down his decisions, and makes it necessary to use people instead of machines for his work, is the problem of how to combine a number of bits of information so that the best selection is made, the best advice is given, or the most effective therapy is planned.

Testmakers, recognizing this problem, try to make things easier by providing total scores, whenever they can properly do so. When a single test is given, the outcome is generally expressed as a total of the scores on the items. When several tests are administered, the results often are combined into some type of sum of scores. In effect, a total score is a summary of a number of observations of behavior.

We compute a total score for just one solid reason: we believe it will give us a better assessment of what is being measured than any of its parts. The skeptical user may ask, "Is there the right amount of each part in the total, for my purposes?" In more psychometric terms, the question refers to the *weight* of each part in the total. It is, of course, obvious that the parts contribute to the total. How much each part contributes is not always obvious. Since the effectiveness of a total score may depend on the weights of its parts, test users might well consider how such weights operate.

Self-Weighting

The most elementary type of total score is the simple summation. When several different tests have been given, this total would be the sum of the raw scores on the tests. Contrary to a widespread belief, such a summation does not automatically yield a total in which the parts are equally weighted, even when all parts contain the same number of items. A simple summation cannot be properly described as "unweighted." As Hull[1] pointed out many years ago, "If tests are not weighted, they weight themselves." In this connection, it may be noted that the number of items and the mean score, of themselves, do not affect the weight of a part in the total. Let us consider, then, some of the factors which determine the weight of a part in a total score.

One method of appraising the relative contributions of parts to a total is to compare the correlations of the parts with the total. If each part has the same correlation with the total we could say that all parts have the same weight in determining the total score. If a total is the sum of scores on Tests A, B, and C, and the correlation between Test A and the total is higher than that found between either of the other tests and the total, then Test A has the greatest weight of the three in the total. The

[1] Hull, C. L. *Aptitude Testing*. Yonkers-on-Hudson, N. Y.: World Book Company, 1928.

The contents of this Bulletin are not copyrighted; the articles may be quoted or reprinted without formality other than the customary acknowledgment of the Test Service Bulletin of THE PSYCHOLOGICAL CORPORATION *as the source.*

correlation between a part and the total, in turn, depends on the standard deviation of the part and the correlations between the part and each of the other parts. Often we do not know the intercorrelations among the parts for the group in which we are interested. We may know these for one or more samples but usually not for a particular group. For practical purposes, we can get a general idea of the contribution of a part by simply comparing its standard deviation with that of every other part. The one with the largest standard deviation ordinarily contributes most to the total and consequently has the greatest actual weight. As noted earlier, the weights of tests are not proportional to their mean scores. A long test, with a high mean score, may have a smaller standard deviation than a much shorter test, but the shorter test would contribute more to the total.

To come back to our total score as a simple summation, we may find that one part has much greater weight in the total than do the other parts. Consider, for example, an office clerical test which consists of vocabulary, arithmetic, and checking, and also yields a total score obtained by summing the raw scores on the parts. For the norms group presented in the test manual, the standard deviations of the parts are given as 8.0 for vocabulary, 4.4 for arithmetic, and 13.9 for checking. Clearly, the checking part has the greatest weight in the total score although the total is called an "unweighted total."

How does this affect the test user? Often a test is selected because the user feels its parts measure areas which he considers important. A total score may impress him as desirable because it appears to be an over-all measure which gives about equal weight to each part. He should be aware that the simple total may actually give much greater weight to one of the parts than he would like to have it give. If the standard deviations of the parts are shown in the manual, an estimate of the relative contribution of each part to the total can readily be determined.

Altering Weights

Since the size of the standard deviation is a crucial matter in weighting, we could bring about changes in weights by altering the standard deviations. If, for example, we want equal weights for all tests or parts which enter into a total, we could transform the scores on every test so that the standard deviations would be the same. Such transformed scores are generally called scaled or standard scores. Neglecting variations which arise from differences in intercorrelations among the tests, a summation of the scaled scores would yield a total in which the separate tests are equally weighted.

Numerous methods are available for transforming scores. Usually, the mean and standard deviation of each test are converted to predetermined numbers. Making all means equal to each other is simply a convenience; it is not necessary for achieving equal weights for the tests. As an example, consider the *Wechsler Adult Intelligence Scale (WAIS)*. The author believed that each of the eleven subtests was equally indicative of intelligence and therefore should be given equal weight in an over-all measure of IQ. To add the raw scores on the tests would have meant that each test contributed to the total in proportion to its standard deviation. In the *WAIS*, therefore, the scores on the subtests are transformed to scaled scores, with a mean of 10 and a standard deviation of 3. These scaled scores are then summed to give a Verbal Score, a Performance Score, and a Full Scale Score.

For any group of tests the author may assign weights which, in his judgment, are most appropriate. Usually such weighting requires an extra step in the scoring process—that of transforming the scores on each test to another scale. If the author can then demonstrate, by means of data, that his decision on weights was good, the user will be much happier about the extra step.

We have considered totals which are simple summations of raw scores and totals in which each test or part is equally weighted. Both types are, in a sense, totals based on judgment. That is to say, the author, consciously or otherwise, determined the weight of each part in the total. When he decided to use a simple sum of raw scores, he was using weights for the parts (or tests) which were proportional to the standard deviations; when he decided to use equal weights, the author again based the combination on his best judgment. Although we would not condemn an author's judgment in these matters, it is reassuring to know there are ways in which such judgment can be verified. When relevant criteria are available, we can determine the effectiveness of different types of totals by correlating each with the criterion. Fortunately, we do not have to compute a large variety of totals to find the most predictive type.

Most Predictive Weights

When several test scores can be used for predicting a criterion, it is possible to weight each test so that the resulting total score has the highest possible correlation with the criterion. This is done by using the *methods of multiple regression* to determine the weights. A total computed in this way can never be a poorer predictor than any of its parts. At worst, it will be as good as the

most predictive part; at best, it will be much more predictive than any part. If our goal is to predict a particular criterion from a number of different tests, or parts of a test, the ideal procedure is to determine the best weight for each test and use the sum of the weighted scores. Unfortunately, weights which are best for one criterion need not be, and generally are not, best for another criterion. Consider, for example, data derived from a study of the *Differential Aptitude Tests* in relation to the prediction of scores on the College Entrance Examination Board tests.

The *DAT* battery yields eight scores. Verbal Reasoning, Numerical Ability, Abstract Reasoning, Space Relations, Mechanical Reasoning, Clerical Speed and Accuracy, Spelling, and Sentences. The authors of the battery do not recommend any type of total score for general use because the primary purpose of the battery is to provide information about strengths and weaknesses in each of the areas measured. One of the purposes of the study was to determine how well *DAT* scores obtained early in Grade 10 predicted scores on the Verbal and Numerical parts of the *CEEB Scholastic Aptitude Test* administered late in Grade 12. It was found that an optimally weighted combination[2] of three tests—Verbal Reasoning (V), Spelling (Sp), and Sentences (Se)—was correlated to the extent of .79, for each sex, with *SAT*-Verbal scores. The total score was obtained by the formula 7V + Sp + Se. The Numerical part of the *SAT* was predicted with even greater accuracy by a combination of scores on Numerical Ability (N), Verbal Reasoning (V), and Space Relations (S). When the total was computed according to the formula 7N + 4V + S, the resulting coefficients with *SAT*-N were .85 for boys and girls.[3] It may be noted that only three of the eight tests were used to predict each criterion, that only one test was common to the two types of total score, and even for the common test, the weight did not remain constant.

Although such best or optimum weights will yield the total which is most highly correlated with a specific criterion, there are certain practical difficulties in general use of the multiple regression procedure. When we use a multi-purpose battery such as the *Differential Aptitude Tests*, we are bound to consider the prediction of many types of criteria. Consequently, a large number of sets of weights would be needed. This is likely to become somewhat confusing, not to mention very tedious, as we change weights for each criterion and multiply a score by its weight and add it to another product of score and weight, etc. Indeed, one of the arguments for a total which is the simple summation of raw scores is that it is easy to compute. However, we must remember that a total so obtained may be much less efficient as a predictor than an optimally weighted total, or even, in some cases, poorer than the best subtest alone. The only way to find out is to compare the correlations between each criterion and the two types of totals.

Planning for a Total Score

Occasionally a battery is constructed for which the addition of raw scores yields a total which closely approximates the best-weighted combination for predicting a desired criterion. This type of total embodies the ease of computation inherent in the simple addition of raw scores and the maximum predictive power inherent in the use of optimum weights. The *College Qualification Tests (CQT)* are an example of such a battery. The purpose of these tests is to serve colleges in their admission, placement, and guidance procedures. The *CQT* include a Verbal test, a Numerical test, and an Information test composed of items from the fields of science and social studies. For a battery intended to select college freshmen, it is obvious that one of the most important and common criteria to be predicted would be the over-all freshman average or "grade point average." The test authors therefore wanted a total score which would predict grade point average as effectively as possible. This, of course, involved optimal weighting of the three test scores. There was, however, another goal—to approximate the best weighting by simply summing the raw scores.

To achieve these ends, the authors of the *CQT* had to estimate the standard deviation needed for each test so that the sum of raw scores would be optimally weighted for predicting freshman grade point average. One cannot, of course, make very precise estimates of such standard deviations. Students' scores and grade point averages vary considerably from college to college. The optimum combination of tests for one institution may be considerably less than optimum for another school. However, the *CQT* authors concluded from their review of available studies that the best combination would be approximated if the standard deviation of the Verbal scores was about 1½ times the standard deviations of the Numerical or Information scores. The three tests were constructed with this goal in mind. The results of

[2]Actually, the optimum weights were rounded to single-digit numbers to facilitate computation.

[3]Seashore, H. G. Tenth grade tests as predictors of twelfth grade scholarship and college entrance status. *J. counsel. Psychol.*, 1954, **1**, 106-115. A reprint of this article, giving the complete prediction formulas, will be sent free on request.

the standardization testing indicated that the desired ratios of standard deviations had been achieved. There remained the test of the hypothesis that the simple summation of raw scores was approximately equivalent to the best-weighted combination.

Data relevant to this problem were presented[4] by two of the authors of the *CQT*. For each of four colleges where *CQT* test scores and freshman grade point averages were available, two types of correlation coefficients were obtained. The multiple coefficient of correlation was computed between the three *College Qualification Tests* and the criterion to indicate relationship between the tests, *when optimally weighted,* and the grade point average. The coefficient of correlation between the *simple sum of raw scores* (*CQT* Total) and the criterion of grade point average was also computed. The data are shown in Table 1. It may be seen that there is very little difference between the correlation coefficients obtained from optimally weighted total scores and from the simple addition of raw scores. In other words, the standard deviations of the tests were such that the simple addition of scores resulted in a close approximation of the best weighting of the tests for each institution.

We cannot, of course, expect to find that the *CQT* Total is optimally weighted for the prediction of specific course grades as well as for over-all grade point average. Studies of this type need to be made. It is, however, encouraging to find that tests can be constructed so that a simple summation of raw scores approximates optimum weighting for predicting a principal criterion.

* * *

In summary, let us consider the total score concept from the test user's point of view. When he selects a group of tests he does so because he wishes to predict certain criteria. It seems reasonable that a total of scores on the various tests would aid in achieving this goal. Obviously the simplest kind of total is the one obtained by adding raw scores on the tests. This type of total, however, automatically assigns weights to the tests which are proportional to their standard deviations. By inspecting the standard deviations of the tests, the user can estimate the relative contribution of each test to the total score. Such inspection can sometimes be very revealing.

When desired or when it seems appropriate to do so, tests can be weighted equally in a total by transforming the raw scores on each test so that the standard deviations will be the same. Usually, this is done by converting the scores on each test to scaled scores. This procedure does not, of itself, provide a better measuring instrument. We still have to conduct validation studies.

For a particular criterion, the most predictive type of total score is one based on optimum weights for the tests. This usually requires additional work on the part of the user. In some instances, test authors have so constructed their tests that a simple summation of scores is also the best-weighted total for a specific type of criterion. But in most cases, it is not safe to assume that a simple combination of raw scores is the best total to use. The data must be examined with care, or we may find ourselves with weight where we least want it. – J.E.D.

Table 1. Comparison of Validity Coefficients Derived from *CQT* Total Score and from Regressed Weight Sum of Three *CQT* Scores

Institution	Sex	N	Validity Coefficients	
			From CQT Total	From Regressed Weight Sum
A	M	449	.46	.46
	F	262	.59	.59
B	M	151	.51	.54
	F	169	.65	.68
C	M	217	.60	.60
	F	76	.52	.56
D	F	107	.71	.71

[4]Wesman, A. G., & Bennett, G. K. Multiple regression vs. simple addition of scores in prediction of college grades. *Amer. Psychologist,* 1957, **12,** 409 (Abstract).